INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Burst Reduction Properties of Rate-based Flow Control Schemes: Downstream Queue Behavior

Zhen LIU, Don TOWSLEY

N° 2117
Octobre 1993

PROGRAMME 1

Architectures parallèles,
bases de données,
réseaux et systèmes distribués

Burst Reduction Properties of Rate-Based Flow Control Schemes: Downstream Queue Behavior

Zhen LIU*
INRIA Centre Sophia Antipolis
2004 Route des Lucioles
06560 Valbonne
France

Don TOWSLEY†
Dept. of Computer Science
University of Massachusetts
Amherst, MA 01003
U.S.A.

October 1993

Abstract

In this paper we consider rate-based flow control throttles feeding a sequence of single server infinite capacity queues. Specifically, we consider two types of throttles, the *token bank* and the *leaky bucket*. We show that the cell waiting times at the downstream queues are increasing functions of the token buffer capacity. These results are established when the rate-based throttles have finite capacity data buffers as well as infinite capacity buffers. In the case that the data buffer has finite capacity, we require that the sum of the capacities of the data buffer and token buffer be a constant. Last, we establish similar results for the process of number of losses at the last downstream queue in the case that the waiting buffer has finite capacity.

Keywords: ATM, leaky bucket, rate-based flow control, token bank.

*The work of this author was supported in part by the CEC DG XIII under the ESPRIT BRA grant QMIPS.
†The work of this author was supported in part by the National Science Foundation under grants ASC-8802764 and NCR-911618.

Propriétés de lissage de trafic des mécanismes de contrôle de taux : comportements des files d'attente en aval

Résumé

Nous étudions des mécanismes de contrôle de taux qui alimentent un système de files d'attente en tandem. Nous considérons deux types de mécanismes de contrôle de taux : *token bank* et *leaky bucket*. Nous démontrons que les temps d'attentes des cellules dans les files d'attente en aval sont des fonctions croissantes de la capacité du tampons de jetons et de la fréquence de génération de jetons. Ces résultats ont été obtenus quand ces mécanismes de contrôle ont des tampons de données de capacité aussi bien finie qu'infinie. Nous établissons des résultats similaires pour le processus des nombres de pertes dans la dernière file d'attente si celle-ci a un tampon de capacité finie.

Mots-clés: Réseaux ATM, contrôle de flue, leaky bucket, token bank, temps d'attente, nombre de pertes.

1 Introduction

Rate-based flow control has been proposed as a mechanism for reducing the burstiness of traffic sources in high-speed networks (e.g., ATM). One mechanism that has received considerable attention is the *leaky bucket rate-based flow control throttle* [16]. Numerous performance studies have evaluated the effectiveness of this mechanism through either simulation or analysis (see [15, 14, 2] for examples of such studies). The goal of this paper is to investigate *qualitatively* the burst reduction properties of two variations of this mechanism: the *token bank* and the *leaky bucket*. Specifically, we study the effect that different parameters have on waiting times, queue lengths, and losses at switches handling traffic allowed to enter the network from such a throttle.

A rate-control throttle is associated with each source. The throttle generates tokens periodically and each cell generated by a source is required to pair up with a token before it is allowed to enter the network. In the event that no token is available at the time that a cell arrives, the cell is stored in a data buffer. As soon as a token is generated, the oldest cell is permitted to enter the network. Similarly, there is a token buffer associated with the token generator. Cells entering the network then traverse a sequence of switches until they reach their destination. The token bank and leaky bucket differ from each other in their behavior when the token buffer is full. In this case, the token bank continues to generate tokens (which are thrown out so long as the token buffer is full) whereas the leaky bucket ceases generating tokens until the token buffer becomes non-full.

In this paper we examine the effect of the token generation rate and token buffer capacity on the delays that cells incur while going through individual switches of the network. We model the path taken by cells belonging to a single source as a tandem queueing network fed by a traffic source controled by a rate-control throttle. Interfering traffic at each queue in the network is accounted for by increasing the cell service time. Some of our results are based on the assumption that service times are deterministic whereas others are not. Such a model can be used to represent a system that partitions bandwidth between sources at each switching node (e.g., [7, 8, 6]).

Using sample path arguments, we show that the cell delay at each switch decreases as the token buffer capacity decreases and/or the token generation rate decreases when no losses occur. In the case that the buffer corresponding to the last switch on the path has limited

buffer capacity, we show that the number of losses is an increasing function of the token buffer capacity and/or the token generation rate. We also establish comparisons between the token bank and leaky bucket.

Several papers have also studied the burst reduction properties of rate-control throttles. Results on the effects of different parameters on the departure process from the throttle can be found in [9, 10] for the case of the token bank, and [12] for both throttles. Anantharam and Konstantopoulos [1] studied the effect of varying the token buffer capacity on the buffer occupancy of *the first queue* on the path in the network. They showed that the stationary buffer occupancy at this queue with deterministic service time is a stochastically increasing function of the token buffer size for the case of a token bank. Low and Varaiya [13], using a fluid model, obtained several monotonicity properties in the case of a rate-based throttle feeding a tandem queueing network. However, using a fluid model does not allow then to distinguish between the token bank and leaky bucket — both reduce to the same model using their approach.

Last, in a related study, Budka [4] examined monotonicity and convexity properties of the throughput of several rate-control throttles including the token bank and leaky bucket using sample path arguments. Our use of the terminology "token bank" and "leaky bucket" is taken from this study. Berger and Whitt [3] used similar arguments to study the effect of buffer allocation between the token bank data buffer and the buffer at the first downstream node.

The remainder of the paper is organized as follows. Section 2 defines and introduces a formal model for the two throttles feeding a tandem queueing system. Section 3 establishes the equivalence between two LB or TB schemes having different token and data buffer capacities. Section 4 contains preliminary sample path comparison results that will be needed in the paper. Monotonicity results regarding cell delays and queue lengths at each node of the network in the case of no losses are established in Sections 5 and 6 respectively. Monotonicity results regarding losses are shown in Section 7. Finally, concluding remarks are provided in Section 8

2 Model and Notation

We consider a rate-control throttle feeding a tandem queueuing network consisting of J single server queues. The first $J-1$ queues, labelled $1, \ldots, J-1$ have infinite capacity and deterministic service times whereas the last queue, J, has a finite capacity c and independent and identically

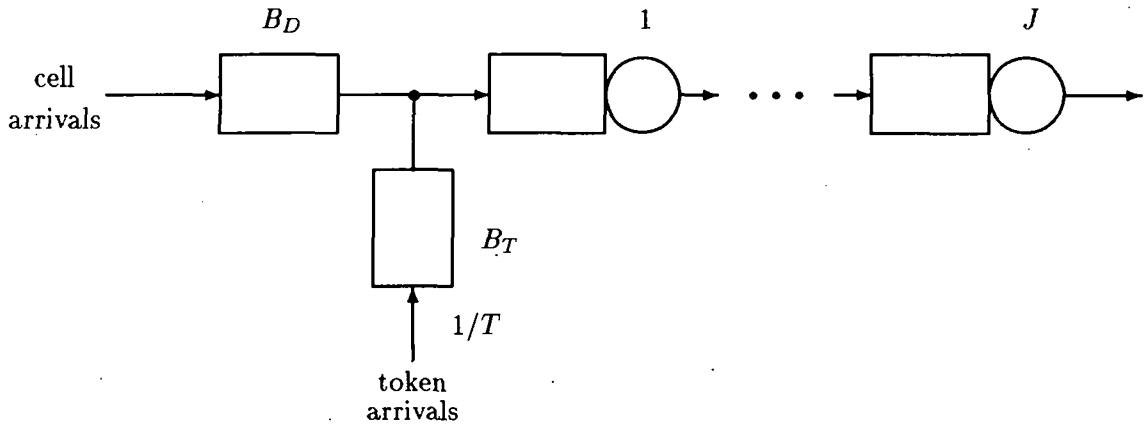

Figure 1: Rate-control throttle feeding a tandem queueing system.

distributed service times taken from an arbitrary distribution (Figure 1). Two rate-control throttles will be considered: the token bank (TB) and the leaky bucket (LB). Both throttles contain a data buffer of size $0 \leq B_D \leq \infty$ and a token buffer of size $0 \leq B_T \leq \infty$. Let $B = B_D + B_T$ be the total buffer size. In order to avoid triviality, we assume that $B \geq 1$.

The two throttles differ slightly in the way the tokens are generated. The token bank generates tokens periodically with constant rate $T^{-1} < \infty$ (or period length $T > 0$). A generated token is accepted by the token buffer if there is empty space, i.e., if the token buffer is not full. A token that finds the token buffer full at the time of its arrival is rejected.

The leaky bucket generates tokens periodically with constant rate $T^{-1} < \infty$. When the token buffer is full, the token generator is shut off. The token generator is turned on again when the token buffer has space for at least one token. Note that, if at time t the queue length of the token buffer drops from B_T to $B_T - 1$, the next token arrival occurs at time $t + T$.

When a cell (or fixed length packet) arrives, it is accepted by the data buffer if it not full. A cell that finds the data buffer full at the time of its arrival is rejected. A cell leaves the data buffer and is transmitted to the downstream system if there is a token in the token buffer. When a cell leaves the data buffer, it consumes one token. i.e., a token leaves the token buffer at the same time. By convention, we will assume that when a token and a cell arrive simultaneously in the system, both the cell and the token are accepted, whatever the status of the data buffer and the token buffer may be. In such a case, a cell and a token leave the system simultaneously.

The leaky bucket is not defined when $B_T = 0$. We follow the convention in this case that

there is a token buffer which is always empty. Under such a convention, the leaky bucket behaves exactly in the same way as the token bank.

Note that there is no need for a buffer to store tokens and that a counter suffices. The token buffer serves to visualize the way these mechanisms work.

We define the following notation concerning the flow control scheme:

- a_n: arrival time of the n-th cell; $a_1 > 0$; for notational simplicity and without loss of generality, we assume that there is at most one cell arrives at any time;
- $\alpha_n = a_n - a_{n-1}$: n-th inter-arrival time; $\alpha_1 = a_1$;
- $\hat{a}_n$: arrival time of the n-th accepted cell; $\hat{a}_1 = a_1$;
- g_n: time epoch when the n-th token is generated; by convention, we assume $g_n = 0$, $1 \leq n \leq B_T$, and $g_{B_T+1} > 0$;
- $\hat{g}_n$: the arrival time of the n-th accepted token; by convention, $\hat{g}_n = 0$, $1 \leq n \leq B_T$, and $\hat{g}_{B_T+1} > 0$;
- d_n: time epoch of the n-th cell departure; It is clear that $d_n = \max(\hat{g}_n, \hat{a}_n)$;
- $\delta_n = d_n - d_{n-1}$: n-th cell inter-departure time; $\delta_1 = d_1$.

We define the following notation. Cells arrive to the data buffer at time $a_1 < a_2 < \cdots$ where $\alpha_n = a_n - a_{n-1}$ denotes the n-th interarrival time, $n = 1, 2, \ldots$, with $a_0 = 0$. Tokens are generated at times $g_1, g_2, \ldots$ with the convention $g_n = 0$, $1 \leq n \leq B_T$, and $g_{B_T+1} > 0$. Let $\hat{a}_n$ and $\hat{g}_n$ denote the arrival time of the n-th accepted cell and token respectively; by convention, $\hat{g}_n = 0$, $1 \leq n \leq B_T$, and $\hat{g}_{B_T+1} > 0$. Let $d_n = \max(\hat{g}_n, \hat{a}_n)$ denote the time of the departure of the n-th cell. Let δ_n denote the n-th cell inter-departure time; $\delta_n = d_n - d_{n-1}$ and $\delta_1 = d_1$.

Let $A = \{a_n\}_{n=1}^{\infty}$ and $G = \{g_n\}_{n=1}^{\infty}$ be the cell arrival and token generation time sequences respecively; $\hat{A} = \{\hat{a}_n\}_{n=1}^{\infty}$ and $\hat{G} = \{\hat{g}_n\}_{n=1}^{\infty}$ be the arrival sequences of accepted cells and tokens respectively. Note that in the leaky bucket scheme, sequences G and $\hat{G}$ are identical. Denote by $V = \{v_n\}_{n=1}^{\infty}$ the indices of accepted cells, viz., $\hat{a}_n = a_{v_n}$. Note that when the data buffer is infinite, the sequences A and $\hat{A}$ coincide, and $v_n = n$ for all $n = 1, 2, \cdots$. Let $\boldsymbol{\alpha} = \{\alpha_n\}_{n=1}^{\infty}$ and $\boldsymbol{\delta} = \{\delta_n\}_{n=1}^{\infty}$ be the sequence of inter-arrval times and the sequence of inter-departure times.

Define K to be the set of indices of accepted cells which are instantaneously transmitted to the downstream system: $K = \{n | n \in I\!N_+,\ \widehat{a}_n = d_n\}$, where $I\!N_+$ is the set of strictly positive integers. Let $\overline{K} = I\!N_+ - K$.

We are interested in the following processes: $\{Q_t^D : t \geq 0\}$, the number of cells waiting in the data buffer, $\{Q_t^T : t \geq 0\}$, the number of tokens waiting in the token buffer, $\{Z_t \stackrel{\text{def}}{=} Q_t^D - Q_t^T : t \geq 0\}$, the difference in the queue lengths, and $\{D_t : t \geq 0\}$, the number of departures. These processes are assumed to be right-continuous. Thus, $Q_{\widehat{a}_n}^D$ and $Q_{\widehat{a}_n}^T$ represent the numbers of cells and tokens, respectively, waiting in the system just after the arrival of the n-th accepted cell. Unless otherwise stated, we will assume throughout this paper that the token buffer is initially full ($Q_0^T = B_T$).

Cells departing the throttle enter a tandem queueing network consisting of J single-server queues labelled $j = 1, \ldots, J$, where the first $J-1$ queues have deterministic service times. We introduce the following notation concerning the downstream tandem network:

- σ_n^j: the service time of the n-th cell that arrives to the j-th queue in the downstream network; for $j = 1, \cdots, J-1$, $\sigma_n^j \equiv \sigma^j$ for all $n \geq 1$;
- d_n^j: the n-th departure time from the j-th downstream queue (and the n-th arrival time at the $j+1$-st downstream queue if $j < J$); by convention, $d_n^0 \stackrel{\text{def}}{=} d_n$ denotes the n-th departure time from the control scheme (and the n-th arrival time in the tandem queueing network);
- W_n^j: the waiting time of the n-th cell that arrives to the j-th downstream queue;
- N_n^j: the length of the j-th downstream queue as seen by the n-th cell to arrive to it;
- M_t^j: the length of the j-th downstream queue at time $t \geq 0$.
- U_t^j: the remaining service time of the cell under service (if any) at time $t \geq 0$ in queue j;
- L_t: the number of cells that are lost in the last downstream queue by time $t \geq 0$.

It is understood that $L_0 = M_0^j = U_0^j = 0$. The processes L_t, M_t^j and U_t^j are also assumed to

be right-continuous. Thus, L_{d_n} represent the numbers of losses in the last downstream queue just after the arrival of the n-th cell in that queue.

The above quantities will be parameterized, when necessary, by (1) the type of control scheme, TB (for token bank) or LB (for leaky bucket), (2) the size of the data buffer, (3) the size of the token buffer, and (4) the length of the token generation period. For example, $W_n^j(TB, B_D, B_T, T)$ (resp. $W_n^j(LB, B_D, B_T, T)$) denotes the waiting time of the n-th cell in the j-th downstream queue departing from the token bank (resp. leaky bucket) scheme with data buffer size B_D, token buffer size B_T and token generation period length T.

The reader should note that the service times within the tandem network are not affected by changes in the parameters of the throttle. This might be reasonable in a network in which bandwidth is partitioned between different sessions as exemplified by hierarchical round robin [8], stop-and-go [7], weighted fair queueing [6]. A similar approach has been taken in [13, 5].

3 Duality

The following lemma is a restatement of Theorem 3.1 in [3].

Lemma 3.1 *Consider two rate-control throttles C and $\tilde{C}$ with data buffer sizes $B_D \geq 0$ and $\tilde{B}_D \geq 0$, respectively, and token buffer sizes $B_T \geq 0$ and $\tilde{B}_T \geq 0$, respectively. Assume that $B_D + B_T = \tilde{B}_D + \tilde{B}_T$. Let Q_t^D and Q_t^T (resp. $\tilde{Q}_t^D$ and $\tilde{Q}_t^T$) be the lengths of data buffer and token buffer of scheme C (resp. C') at time t, respectively. If*

$$Q_0^D + (B_T - Q_0^T) = \tilde{Q}_0^D + (\tilde{B}_T - \tilde{Q}_0^T), \tag{3.1}$$

then for any arbitrarily fixed cell arrival sequence $A = \{a_n\}_{n=1}^{\infty}$ and token generation sequence $G = \{g_n\}_{n=1}^{\infty}$, cells (resp. tokens) are accepted in C if and only if they are accepted in C', and

$$Q_t^D + (B_T - Q_t^T) = \tilde{Q}_t^D + (\tilde{B}_T - \tilde{Q}_t^T), \qquad t \geq 0; \tag{3.2}$$

$$(B_D - Q_t^D) + Q_t^T = (B_D - \tilde{Q}_t^D) + \tilde{Q}_t^T, \qquad t \geq 0. \tag{3.3}$$

This lemma has as its direct consequence the following theorem, also established in [3],

Theorem 3.1 *Consider two flow control schemes C and C', that are either both token banks or leaky buckets, with data buffer sizes $B_D \geq 0$ and $B'_D \geq 0$, respectively, and token buffer sizes $B_T \geq 0$ and $B'_T \geq 0$, respectively. If $B_D + B_T = B'_D + B'_T$, then for any token generation period length T and any arbitrarily fixed cell arrival sequence $A = \{a_n\}_{n=1}^{\infty}$, the arrival sequences of accepted cells are identical:*

$$\hat{A}(TB, B_D, B_T, T) = \hat{A}(TB, B'_D, B'_T, T), \tag{3.4}$$

$$\hat{A}(LB, B_D, B_T, T) = \hat{A}(LB, B'_D, B'_T, T). \tag{3.5}$$

Proof. The result for token banks is a result of lemma 3.1. Consider the case of two leaky buckets. According to our convention,

$$\begin{aligned} Q_0^D(LB, B_D, B_T, T) &= Q_0^D(LB, B'_D, B'_T, T) = 0; \\ B_T - Q_0^T(LB, B_D, B_T, T) &= B'_T - Q_0^T(LB, B'_D, B'_T, T) = 0. \end{aligned}$$

Note that if for some t, $Q_t^T(LB, B_D, B_T, T) = B_T \geq 1$ and if

$$\begin{aligned} & Q_t^T(LB, B_D, B_T, T) + B_T - Q_t^T(LB, B_D, B_T, T) \\ &= Q_t^T(LB, B'_D, B'_T, T) + B'_T - Q_t^T(LB, B'_D, B'_T, T), \end{aligned}$$

then $Q_t^D(LB, B_D, B_T, T) = 0$, so that $Q_t^T(LB, B'_D, B'_T, T) = 0$ and $Q_t^T(LB, B'_D, B'_T, T) = B'_T$. Therefore, it is readily proved by induction (as in Lemma 3.1) that at event times $0 = t_1 < t_2 < \cdots < t_n < \cdots$, a cell (resp. token) is accepted at time t_n in C if and only if they are accepted in C', and that the token buffer is full in C if and only if it is full in C'.

Thus, the token arrival times are identical in both schemes, so that Lemma 3.1 still applies. ∎

Remark: Owing to Lemma 3.1, the duality holds for general token arrival sequence and arbitrary initial token number provided the initial condition (3.1) is satisfied.

4 Basic Sample Path Characteristics

We now present several preliminary comparison relations which will be used to prove our main results.

The first one characterizes the departure process of cells from the throttle.

Theorem 4.1 *Let C be a rate-control throttle with token generation period length T, token buffer size $B_T \geq 1$ and the arrival sequence of accepted cells $\hat{A}$. Let $K = \{n_1, n_2, \cdots, n_k, \cdots, n_{k_0}\}$, where $k_0 \leq \infty$, $1 = n_1 < n_2 < \cdots < n_k < \cdots < n_{k_0}$ (by convention, $n_{k_0} = \infty$ if $k_0 = \infty$). Then, for all $k \geq 2$,*

$$d_{n_k} - d_{n_{k-1}} = \hat{a}_{n_k} - \hat{a}_{n_{k-1}}. \tag{4.1}$$

Moreover, for all $2 \leq k \leq k_0$, if $n_k > n_{k-1} + 1$, then

$$\delta_{n_{k-1}+1} \leq T, \quad \delta_{n_{k-1}+2} = \cdots = \delta_{n_k - 1} = T, \quad \delta_{n_k} \geq T. \tag{4.2}$$

Further, if $k_0 < \infty$, then

$$\delta_{n_{k_0}+1} \leq T, \qquad \delta_i = T, \quad i \geq n_{k_0} + 2. \tag{4.3}$$

Proof. Equation (4.1) trivially follows from the definition of the set of instantaneous departure points K.

Assume now $n_k > n_{k-1} + 1$. Observe that if the data buffer size is zero, then $n_k = k$ for all $k \geq 0$. Therefore, relation $n_k > n_{k-1} + 1$ implies that the data buffer size is at least one. Consider i such that $n_{k-1} + 1 \leq i \leq n_k - 1$. Now $i \notin K$ so that $\hat{a}_i < d_i = \hat{g}_i$. As this is true for all i, $n_{k-1} + 1 \leq i \leq n_k - 1$, the token buffer is always empty during the time interval $(d_{n_{k-1}}, d_{n_k - 1}]$. Therefore, as $B_T \geq 1$, no token is lost during the time interval $(d_{n_{k-1}}, d_{n_k - 1}]$. Thus, for all $n_{k-1} + 2 \leq i \leq n_k - 1$

$$\delta_i = d_i - d_{i-1} = \hat{g}_i - \hat{g}_{i-1} = T.$$

Moreover,

$$\delta_{n_k} = d_{n_k} - d_{n_k - 1} \geq \hat{g}_{n_k} - d_{n_k - 1} = \hat{g}_{n_k} - \hat{g}_{n_k - 1} = T.$$

Since $n_{k-1}+1 \notin K$, $Q_t^T = 0 < B_T$ for all $d_{n_{k-1}} < t < d_{n_{k-1}+1}$. Thus, no token is generated during the time interval $(d_{n_{k-1}}, d_{n_{k-1}+1})$, so that

$$\delta_{n_{k-1}+1} = d_{n_{k-1}+1} - d_{n_{k-1}} \leq T.$$

Hence, relation (4.2) holds.

Relation (4.3) is shown by the same arguments. ∎

The remainder of this section is devoted to deriving inequalities among the arrival times of accepted tokens for different parameters. This is based on the following evolution equations.

Theorem 4.2 *Assume $B_T \geq 1$. Then for any cell arrival sequence $\boldsymbol{\alpha}$, and for all $n \geq 1$,*

$$\hat{g}_{n+1}(LB, B_D, B_T, T) = \max(\hat{g}_n(LB, B_D, B_T, T),\ \hat{a}_{n+1-B_T}) + T, \tag{4.4}$$

$$\hat{g}_{n+1}(TB, B_D, B_T, T) = \lceil \max(\hat{g}_n(TB, B_D, B_T, T) + T,\ \hat{a}_{n+1-B_T})/T \rceil \cdot T, \tag{4.5}$$

where $\lceil x \rceil$ denotes the smallest integer which is greater than or equal to x. By convention, $\hat{a}_i = 0$ if $i \leq 0$.

Proof. Consider first relation (4.4). According to our convention, $\hat{g}_n(LB, B_D, B_T, T) = 0$ for all $1 \leq n \leq B_T$. Therefore, equation (4.4) holds for $1 \leq n \leq B_T - 1$.

For $n \geq B_T$, if $\hat{g}_n(LB, B_D, B_T, T) < \hat{a}_{n+1-B_T}$, then

$$Q^T_{\hat{g}_n(LB,B_D,B_T,T)}(LB, B_D, B_T, T) = B_T,$$

and the first token departure after time $\hat{g}_n(LB, B_D, B_T, T)$ occurs at time $\hat{a}_{n+1-B_T}$. Hence,

$$\hat{g}_{n+1}(LB, B_D, B_T, T) = \hat{a}_{n+1-B_T} + T = \max(\hat{g}_n(LB, B_D, B_T, T),\ \hat{a}_{n+1-B_T}) + T.$$

If, however, $\hat{g}_n(LB, B_D, B_T, T) \geq \hat{a}_{n+1-B_T}$, then

$$Q^T_{\hat{g}_n(LB,B_D,B_T,T)}(LB, B_D, B_T, T) < B_T,$$

so that

$$\hat{g}_{n+1}(LB, B_D, B_T, T) = \hat{g}_n(LB, B_D, B_T, T) + T = \max(\hat{g}_n(LB, B_D, B_T, T),\ \hat{a}_{n+1-B_T}) + T.$$

In both cases, equation (4.4) holds.

Similarly, for token bank, by our convention, $\hat{g}_n(TB, B_D, B_T, T) = 0$ for all $1 \leq n \leq B_T$. Therefore, equation (4.5) holds for $1 \leq n \leq B_T - 1$.

For $n \geq B_T$, if $\hat{g}_n(TB, B_D, B_T, T) < \hat{a}_{n+1-B_T}$, then

$$Q^T_{\hat{g}_n(TB,B_D,B_T,T)}(TB, B_D, B_T, T) = B_T,$$

and that the first token departure after time $\hat{g}_n(TB, B_D, B_T, T)$ occurs at time $\hat{a}_{n+1-B_T}$. Hence,

$$\hat{g}_{n+1}(TB, B_D, B_T, T) = \left\lceil \frac{\hat{a}_{n+1-B_T}}{T} \right\rceil \cdot T = \left\lceil \frac{\max(\hat{g}_n(TB, B_D, B_T, T) + T,\ \hat{a}_{n+1-B_T})}{T} \right\rceil \cdot T.$$

If, however, $\hat{g}_n(TB, B_D, B_T, T) \geq \hat{a}_{n+1-B_T}$, then

$$Q^T_{\hat{g}_n(TB,B_D,B_T,T)}(TB, B_D, B_T, T) < B_T,$$

so that

$$\hat{g}_{n+1}(TB, B_D, B_T, T) = \hat{g}_n(TB, B_D, B_T, T) + T = \left\lceil \frac{\max(\hat{g}_n(TB, B_D, B_T, T) + T,\ \hat{a}_{n+1-B_T})}{T} \right\rceil \cdot T.$$

In both cases, the equation (4.5) holds. ∎

Corollary 4.1 *Assume $B_T \geq 1$. Then for any cell arrival sequence α, and for all $n \geq 1$,*

$$\hat{g}_{n+1}(LB, \infty, B_T + 1, T) = \hat{g}_n(LB, \infty, B_T, T) \quad (4.6)$$

$$\hat{g}_{n+1}(TB, \infty, B_T + 1, T) = \hat{g}_n(TB, \infty, B_T, T) \quad (4.7)$$

Proof. Relations (4.6) and (4.7) can simply be shown by induction on n using the evolution equations (4.4) and (4.5). The detailed proof is omitted. ∎

Theorem 4.3 *Assume $B_T \geq 1$. Then for any cell arrival sequence α,*

$$\hat{g}_n(TB, \infty, B_T, T) \leq \hat{g}_n(LB, \infty, B_T, T) \quad (4.8)$$

$$\widehat{g}_n(LB,\infty,B_T+1,T) \leq \widehat{g}_n(TB,\infty,B_T,T) \tag{4.9}$$

$$\widehat{g}_n(LB,\infty,B_T,T) \leq \widehat{g}_n(LB,\infty,B_T,T'), \qquad T' \geq T \tag{4.10}$$

$$\widehat{g}_n(TB,\infty,B_T,T) \leq \widehat{g}_n(TB,\infty,B_T,T'), \qquad T' = mT, \quad m \in I\!N_+ \tag{4.11}$$

$$\widehat{g}_n(TB,\infty,B_T+1,T) \leq \widehat{g}_n(TB,\infty,B_T,T'), \qquad T' \geq T \tag{4.12}$$

Proof. We will prove these inequalities by induction on n. Clearly, all these relations are true for $1 \leq n \leq B_T$. Assume they hold for some $n \geq B_T$.

It then follows from the evolution equation (4.5) that

$$\begin{aligned}
\widehat{g}_{n+1}(TB,\infty,B_T,T) &\leq \max(\widehat{g}_n(TB,\infty,B_T,T),\widehat{a}_{n+1-B_T}) + T \\
&\leq \max(\widehat{g}_n(LB,\infty,B_T,T),\widehat{a}_{n+1-B_T}) + T \\
&= \widehat{g}_{n+1}(LB,\infty,B_T,T),
\end{aligned}$$

where the second inequality comes from the inductive assumption, and the equality from (4.4). Therefore, by induction, relation (4.8) holds for all $n \geq 1$.

In order to show (4.9), we first prove the following inequality:

$$\widehat{g}_n(LB,\infty,B_T,T) \leq \widehat{g}_n(TB,\infty,B_T,T) + T, \qquad n \geq 1. \tag{4.13}$$

Clearly, (4.13) holds for all $1 \leq n \leq B_T$. Assume it holds for some $n \geq B_T$. Applying (4.5) entails that

$$\begin{aligned}
\widehat{g}_{n+1}(LB,\infty,B_T,T) &= \max(\widehat{g}_n(LB,\infty,B_T,T),\widehat{a}_{n+1-B_T}) + T \\
&\leq \max(\widehat{g}_n(TB,\infty,B_T,T) + T,\widehat{a}_{n+1-B_T}) + T,
\end{aligned}$$

where the inequality comes from the inductive assumption. Using further (4.5) implies that

$$\begin{aligned}
\widehat{g}_{n+1}(LB,\infty,B_T,T) &\leq \max(\widehat{g}_n(TB,\infty,B_T,T) + T,\widehat{a}_{n+1-B_T}) + T \\
&\leq \widehat{g}_{n+1}(TB,\infty,B_T,T) + T.
\end{aligned}$$

Therefore, relation (4.13) holds for all $n \geq 1$.

Relations (4.6) and (4.13) imply that

$$\begin{aligned}
\widehat{g}_{n+1}(LB,\infty,B_T+1,T) &= \widehat{g}_n(LB,\infty,B_T,T) \\
&\leq \widehat{g}_n(TB,\infty,B_T,T) + T \\
&\leq \widehat{g}_{n+1}(TB,\infty,B_T,T).
\end{aligned}$$

Thus, by induction, inequality (4.9) holds for all $n \geq 1$.

Consider now relation (4.11). It follows from (4.5) that

$$\begin{aligned}
\hat{g}_{n+1}(TB, \infty, B_T, T') &= \left\lceil \frac{\max(\hat{g}_n(TB, \infty, B_T, T') + T',\ \hat{a}_{n+1-B_T})}{T'} \right\rceil \cdot T' \\
&\geq \left\lceil \frac{\max(\hat{g}_n(TB, \infty, B_T, T') + T',\ \hat{a}_{n+1-B_T})}{T} \right\rceil \cdot T \\
&\geq \left\lceil \frac{\max(\hat{g}_n(TB, \infty, B_T, T) + T,\ \hat{a}_{n+1-B_T})}{T} \right\rceil \cdot T \\
&= \hat{g}_{n+1}(TB, \infty, B_T, T),
\end{aligned}$$

where the first inequality uses the fact T' is an integer multiple of T, the second inequality comes from the inductive assumption. Hence, (4.11) holds for all $n \geq 1$.

In order to establish relation (4.12), we first show the following inequality:

$$\hat{g}_n(TB, \infty, B_T, T) \leq \hat{g}_n(TB, \infty, B_T, T') + T'. \qquad n \geq 1. \tag{4.14}$$

Clearly, (4.14) holds for all $1 \leq n \leq B_T$. For $n \geq B_T$,

$$\begin{aligned}
\hat{g}_{n+1}(TB, \infty, B_T, T) &\leq \max(\hat{g}_n(TB, \infty, B_T, T),\ \hat{a}_{n+1-B_T}) + T \\
&\leq \max(\hat{g}_n(TB, \infty, B_T, T') + T',\ \hat{a}_{n+1-B_T}) + T \\
&\leq \left\lceil \frac{\max(\hat{g}_n(TB, \infty, B_T, T') + T',\ \hat{a}_{n+1-B_T})}{T'} \right\rceil \cdot T' + T' \\
&= \hat{g}_{n+1}(TB, \infty, B_T, T')
\end{aligned}$$

Therefore, (4.14) holds for all $n \geq 1$.

Using (4.7) and (4.14) we obtain

$$\begin{aligned}
\hat{g}_{n+1}(TB, \infty, B_T + 1, T) &= \hat{g}_n(TB, \infty, B_T, T) \\
&\leq \hat{g}_n(TB, \infty, B_T, T') + T' \\
&\leq \hat{g}_{n+1}(TB, \infty, B_T, T').
\end{aligned}$$

Thus, by induction, inequality (4.12) holds for all $n \geq 1$.

Finally, we turn to (4.10). It readily follows from (4.4) that

$$\begin{aligned}\widehat{g}_{n+1}(LB,\infty,B_T,T) &= \max(\widehat{g}_n(LB,\infty,B_T,T),\widehat{a}_{n+1-B_T})+T\\ &\le \max(\widehat{g}_n(LB,\infty,B_T,T'),\widehat{a}_{n+1-B_T})+T'\\ &= \widehat{g}_{n+1}(LB,\infty,B_T,T'),\end{aligned}$$

where the inequality comes from the inductive assumption. Therefore, by induction, inequality (4.10) holds for all $n \ge 1$. ∎

Finally, the following property will also be used.

Theorem 4.4 *Consider two rate-control throttles C and C', which have the same arrival sequence of accepted cells $\widehat{A}$, but different arrival sequences of accepted tokens $\widehat{S} = \{\widehat{g}_n\}_{n=1}^{\infty}$ and $\widehat{S}' = \{\widehat{g}'_n\}_{n=1}^{\infty}$, respectively. Let K and K' be the sets of instantaneous departure points of C and C', respectively. Let d_n (resp. d'_n) be the n-th departure time in C (resp. C'). If for all $n \ge 1$, $\widehat{g}_n \ge \widehat{g}'_n$, then $K \subseteq K'$ and $d_n \ge d'_n$.*

Proof. Note that $n \in K$ if and only if $\widehat{a}_n \ge \widehat{g}_n$. Thus, for all $n \in K$, $\widehat{a}_n \ge \widehat{g}_n \ge \widehat{g}'_n$, so that $n \in K'$. Therefore, $K \subseteq K'$. The inequality $d_n \ge d'_n$ follows from the fact that for all $n = 1, 2, \cdots$,

$$d_n = \max(\widehat{a}_n, \widehat{g}_n) \ge \max(\widehat{a}_n, \widehat{g}'_n) = d'_n.$$

∎

5 Comparison of Waiting Times

We first derive comparison results between two $\cdot/G/1$ queues with FCFS service disciplines. These will be essential in establishing the comparison results of this section. Let $Q^{(j)}$, $j = 1, 2$, be two such queues with arrival times $\{t_n^{(j)}\}_{n=1}^{\infty}$, service times $\{\sigma_n^{(j)}\}_{n=1}^{\infty}$, departure times $\{d_n^{(j)}\}_{n=1}^{\infty}$, and waiting times $\{W_n^{(j)}\}_{n=1}^{\infty}$, $j = 1, 2$. We will find the following relations useful.

$$W_{n+1}^{(j)} = (W_n^{(j)} + \sigma_n^{(j)} - t_{n+1}^{(j)} + t_n^{(j)})^+, \tag{5.1}$$

$$W_m^{(j)} = \max\left\{\left(W_n + \sum_{i=n}^{m-1}\sigma_i - t_m^{(j)} + t_n^{(j)}\right)^+, \max_{n+1\le u\le m}\left(\sum_{i=u}^{m-1}\sigma_i - t_m^{(j)} + t_u^{(j)}\right)^+\right\},$$

$$m = n+1, n+2, \ldots, \tag{5.2}$$

$$d_{n+1}^{(j)} = (d_n^{(j)}, t_n^{(j)})^+ + \sigma_{n+1}^{(j)}, \tag{5.3}$$

where $j = 1, 2$, $n = 0, 1, \ldots$, and $(x)^+$ denotes $\max(0, x)$.

Assume that for all $n \ge 1$, $t_n^{(1)} \ge t_n^{(2)}$, and $\sigma_n^{(1)} = \sigma_n^{(2)} = \sigma_n \le T$ where T is a constant. Assume further that there is a set of indices $K = \{n_1, n_2, \cdots, n_k, \cdots, n_{k_0}\}$, where $1 = n_1 < n_2 < \cdots < n_k < \cdots < n_{k_0} \le \infty$ (by convention, $n_{k_0} = \infty$ if $k_0 = \infty$), such that

- for all $1 \le k \le k_0$, $t_{n_k}^{(1)} = t_{n_k}^{(2)}$;
- for all $1 \le k < k_0$ such that $n_{k+1} - n_k \ge 2$:

$$t_i^{(1)} - t_{i-1}^{(1)} \begin{cases} \le T, & i = n_k + 1; \\ = T, & n_k + 2 \le i \le n_{k+1} - 1; \\ \ge T, & i = n_{k+1}. \end{cases} \tag{5.4}$$

- if $k_0 < \infty$ then $t_{n_{k_0}+1}^{(1)} - t_{n_{k_0}}^{(1)} \le T$ and $t_i^{(1)} - t_{i-1}^{(1)} = T$ for all $i \ge n_{k_0} + 2$.

Lemma 5.1 *For the above two $\cdot/G/1$ queues $Q^{(1)}$ and $Q^{(2)}$,*

$$W_n^{(1)} \le W_n^{(2)}, \qquad n = 1, 2, \cdots, \tag{5.5}$$

$$d_n^{(1)} \ge d_n^{(2)}, \qquad n \ge 1. \tag{5.6}$$

If in addition service times are deterministic, then there is a set of indices $\widetilde{K} = \{\tilde{n}_1, \tilde{n}_2, \cdots, \tilde{n}_k, \cdots, \tilde{n}_{\tilde{k}_0}\}$, where $1 = \tilde{n}_1 < \tilde{n}_2 < \cdots < \tilde{n}_k < \cdots < \tilde{n}_{\tilde{k}_0} \le \infty$ (by convention, $\tilde{n}_{\tilde{k}_0} = \infty$ if $\tilde{k}_0 = \infty$), such that

$$d_{\tilde{n}_k}^{(1)} = d_{\tilde{n}_k}^{(2)}, \qquad 1 \le k \le \tilde{k}_0, \tag{5.7}$$

and for all $1 \le k < \tilde{k}_0$ such that $\tilde{n}_{k+1} - \tilde{n}_k \ge 2$:

$$d_i^{(1)} - d_{i-1}^{(1)} \begin{cases} \le T, & i = \tilde{n}_k + 1; \\ = T, & \tilde{n}_k + 2 \le i \le \tilde{n}_{k+1} - 1; \\ \ge T, & i = \tilde{n}_{k+1}, \end{cases} \tag{5.8}$$

and if $\tilde{k}_0 < \infty$ *then*

$$d^{(1)}_{\tilde{n}_{\tilde{k}_0}+1} - d^{(1)}_{\tilde{n}_{\tilde{k}_0}} \leq T = d^{(1)}_i - d^{(1)}_{i-1}, \qquad i \geq \tilde{n}_{\tilde{k}_0} + 2. \tag{5.9}$$

Proof. Consider (5.5). Assume first that $k_0 = \infty$. We show by induction on k that for all $k = 1, 2, \cdots$, relation (5.5) holds for all $n \leq n_k$. It is clear that $W^{(1)}_1 = W^{(2)}_1 = 0$. Assume there is some $k \geq 1$ such that (5.5) holds for all $n \leq n_k$.

It then follows that

$$\begin{aligned} W^{(1)}_{n_k+1} &= (W^{(1)}_{n_k} + \sigma_{n_k} - t^{(1)}_{n_k+1} + t^{(1)}_{n_k})^+ \\ &\leq (W^{(2)}_{n_k} + \sigma_{n_k} - t^{(2)}_{n_k+1} + t^{(2)}_{n_k})^+ \\ &= W^{(2)}_{n_k+1}. \end{aligned}$$

The two equalities are due to (5.1) and the inequality follows from the inductive hypothesis, and the relations $t^{(1)}_n = t^{(2)}_n$, $n \in K$ and $t^{(1)}_n \geq t^{(2)}_n$, $n \notin K$.

If $n_{k+1} = n_k + 1$, then relation (5.5) holds for all $n \leq n_{k+1}$. If, however, $n_{k+1} > n_k + 1$, then relation (5.4) allows us to simplify (5.2) in the case of $\mathcal{Q}^{(1)}$ for all $n_k + 2 \leq m \leq n_{k+1}$ to

$$W^{(1)}_m = \left(W^{(1)}_{n_k} + \sum_{i=n_k}^{m-1} \sigma_i - t^{(1)}_m + t^{(1)}_{n_k} \right)^+,$$

so that

$$\begin{aligned} W^{(1)}_m &= \left(W^{(1)}_{n_k} + \sum_{i=n_k}^{m-1} \sigma_i - t^{(1)}_m + t^{(1)}_{n_k} \right)^+ \\ &\leq \left(W^{(1)}_{n_k} + \sum_{i=n_k}^{m-1} \sigma_i - t^{(2)}_m + t^{(2)}_{n_k} \right)^+ \\ &\leq \left(W^{(2)}_{n_k} + \sum_{i=n_k}^{m-1} \sigma_i - t^{(2)}_m + t^{(2)}_{n_k} \right)^+ \end{aligned}$$

$$\leq \max\left\{\left(W_{n_k}^{(2)} + \sum_{i=n_k}^{m-1}\sigma_i - t_m^{(2)} + t_{n_k}^{(2)}\right)^+, \max_{n_k+1\leq u\leq m}\left(\sum_{i=u}^{m-1}\sigma_i - t_m^{(2)} + t_u^{(2)}\right)^+\right\}$$

$$= W_m^{(2)}.$$

Hence, relation (5.5) holds for all $n \leq n_{k+1}$. By induction, it holds for all $n = 1, 2, \cdots$.

If $k_0 < \infty$, then, the above induction shows that relation (5.5) holds for all $n \leq n_{k_0}$. A similar argument establishes (5.5) for all $n \geq n_{k_0}$.

A simple induction on n using relation (5.3) with $d_1^{(1)} = t_1^{(1)} + \sigma_1 = t_1^{(2)} + \sigma_1 = d_1^{(2)}$ implies that relation (5.6) holds.

Consider now relation (5.7). Define

$$\widetilde{K} = K \bigcup \{n_k + j \mid n_k \in K,\ 1 \leq j \leq n_{k+1} - n_k - 1,\ W_{n_k+j}^{(1)} > 0\}.$$

Let $U_t^{(j)}$ be the unfinished work in queue $Q^{(j)}$ $j = 1, 2$ at time t. It follows from relation (5.6), for all $n \geq 1$, $U_{t_n^{(1)}}^{(1)} \geq U_{t_n^{(2)}}^{(2)}$. Therefore, by the definition of set K,

$$W_{n_k}^{(1)} = W_{n_k}^{(2)}, \qquad n_k \in K, \tag{5.10}$$

so that

$$d_{n_k}^{(1)} = d_{n_k}^{(2)}, \qquad n_k \in K. \tag{5.11}$$

Consider an arbitrary k such that $n_{k+1} - n_k \geq 2$. Let

$$j_0 = \min\{1 \leq j \leq n_{k+1} - n_k \mid W_{n_k+j}^{(1)} = 0\},$$

where, by convention, the minimum over an empty set is taken as $n_{k+1} - n_k$. Then, due to the assumption on the arrival sequence $\{t_n^{(1)}\}$, it is readily seen that

$$W_{n_k+j}^{(1)} \begin{cases} > 0, & 1 \leq j < j_0, \\ = 0, & j_0 \leq j < n_{k+1} - n_k. \end{cases} \tag{5.12}$$

Using further relation (5.10) and the fact that $t_n^{(1)} \geq t_n^{(2)}$ for all $n \geq 1$, we obtain that for all $1 \leq j < j_0$, $W_{n_k+j}^{(2)} > 0$. Hence,

$$d_{n_k+j}^{(1)} = d_{n_k}^{(1)} + j \cdot \sigma = d_{n_k}^{(2)} + j \cdot \sigma = d_{n_k+j}^{(2)}, \qquad 1 \leq j < j_0. \tag{5.13}$$

In view of (5.11) and (5.13), relation (5.7) holds.

For all $j_0 \leq j < n_{k+1} - n_k$, relation (5.12) implies that $d_{n_k+j}^{(1)} = t_{n_k+j}^{(1)} + \sigma$, so that

$$d_{n_k+j_0}^{(1)} - d_{n_k+j_0-1}^{(1)} = t_{n_k+j_0}^{(1)} + \sigma - (t_{n_k+j_0-1}^{(1)} + W_{n_k+j_0-1}^{(1)} + \sigma) \leq t_{n_k+j_0}^{(1)} - t_{n_k+j_0-1}^{(1)} \leq T, \tag{5.14}$$

$$d_{n_k+j}^{(1)} - d_{n_k+j-1}^{(1)} = t_{n_k+j}^{(1)} + \sigma - (t_{n_k+j-1}^{(1)} + \sigma) = t_{n_k+j}^{(1)} - t_{n_k+j-1}^{(1)} = T, \quad j_0 < j < n_{k+1} \tag{5.15}$$

$$d_{n_{k+1}}^{(1)} - d_{n_{k+1}-1}^{(1)} = t_{n_{k+1}}^{(1)} + \sigma - (t_{n_{k+1}-1}^{(1)} + \sigma) = t_{n_{k+1}}^{(1)} - t_{n_{k+1}-1}^{(1)} \geq T. \tag{5.16}$$

Relations (5.14), (5.15) and (5.16) entail (5.8).

In case $\tilde{k}_0 < \infty$, relation (5.9) can be verified in an analogous way. The proof is thus completed. ∎

The following comparison results follow from the above lemma.

Theorem 5.1 *Assume $B_T \geq 1$ and that $T \geq \sigma_n^j$ a.s. for all $1 \leq j \leq J$ and all $n \geq 1$. Then for any fixed cell arrival sequence A,*

$$W_n^j(LB, \infty, B_T, T) \leq W_n^j(TB, \infty, B_T, T), \tag{5.17}$$

$$W_n^j(TB, \infty, B_T, T) \leq W_n^j(LB, \infty, B_T + 1, T), \tag{5.18}$$

$$W_n^j(TB, \infty, B_T, T) \leq W_n^j(TB, \infty, B_T, T'), \qquad T' = mT, m \in I\!N_+ \tag{5.19}$$

$$W_n^j(LB, \infty, B_T, T) \leq W_n^j(LB, \infty, B_T, T'), \qquad T \geq T' \tag{5.20}$$

$$W_n^j(TB, \infty, B_T, T) \leq W_n^j(TB, \infty, B_T', T'), \qquad T \geq T', 1 \leq B_T < B_T' \tag{5.21}$$

for all $1 \leq j \leq J$ and all $n \geq 1$

Proof. The proof is by induction. We only consider relation (5.17). The proofs of the remaining relations follows in a similar manner. For $j = 1$, the basic sample path properties of Theorems 4.2, 4.3, 4.4 allow us to apply Lemma 5.1 and to obtain relation (5.17).

Moreover, for $j = 1$, Theorems 4.2, 4.3 and 4.4 and Lemma 5.1 imply the existence of a set

$$K^j = \{n_1^j, n_2^j, \cdots, n_k^j, \cdots, n_{k_0^j}^j\},$$

where $1 = n_1^j < n_2^j < \cdots < n_k^j < \cdots < n_{k_0^j}^j \leq \infty$ (by convention, $n_{k_0^j}^j = \infty$ if $k_0^j = \infty$), such that

$$d_{n_k^j}^j(LB, \infty, B_T, T) = d_{n_k^j}^j(TB, \infty, B_T, T), \qquad 1 \leq k \leq k_0^j, \tag{5.22}$$

and for all $1 \leq k < k_0^j$ such that $n_{k+1}^j - n_k^j \geq 2$:

$$\delta_i^j(LB, \infty, B_T, T) \begin{cases} \leq T, & i = n_k^j + 1; \\ = T, & n_k^j + 2 \leq i \leq n_{k+1}^j - 1; \\ \geq T, & i = n_{k+1}^j, \end{cases} \tag{5.23}$$

and if $k_0^j < \infty$ then

$$\delta_{n_{k_0^j}^j + 1}^j(LB, \infty, B_T, T) \leq T = \delta_i^j(LB, \infty, B_T, T), \qquad i \geq n_{k_0^j}^j + 2. \tag{5.24}$$

Furthermore,

$$d_n^j(LB, \infty, B_T, T) \geq d_n^j(TB, \infty, B_T, T), \qquad n \geq 1. \tag{5.25}$$

Assume relations (5.22), (5.23), (5.24) and (5.25) for some $1 \leq j < J$. Then, an application of Lemma 5.1 implies that relation (5.17) holds for $j+1$. Hence, (5.17) holds for all $1 \leq j \leq J$. ∎

Theorem 5.1 implies the following monotonicity of the waiting times with respect to the token buffer size.

Corollary 5.1 *Assume $B_T \geq 1$ and for all $1 \leq j \leq J$ and all $n \geq 1$, $T \geq \sigma_n^j$ a.s. Then for any fixed cell arrival sequence A,*

$$W_n^j(TB, \infty, B_T, T) \leq W_n^j(TB, \infty, B_T + 1, T), \qquad 1 \leq j \leq J, \quad n = 1, 2, \cdots, \tag{5.26}$$

$$W_n^j(LB, \infty, B_T, T) \leq W_n^j(LB, \infty, B_T + 1, T), \qquad 1 \leq j \leq J, \quad n = 1, 2, \cdots. \tag{5.27}$$

When the token buffer has infinite capacity, the downstream tandem queueing network is fed by the arrival sequence A. Denote by $W_n^j(A)$ the waiting time of the n-th cell in the j-th downstream queue with network arrival sequence A. The following result indicates that the leaky bucket and the token bank flow control schemes reduce the waiting times:

Corollary 5.2 *Assume $B_T \geq 1$ and for all $1 \leq j \leq J$ and all $n \geq 1$, $T \geq \sigma_n^j$ a.s. Then for any fixed cell arrival sequence A,*

$$W_n^j(TB, \infty, B_T, T) \leq W_n^j(A), \quad 1 \leq j \leq J, \quad n = 1, 2, \cdots, \tag{5.28}$$

$$W_n^j(LB, \infty, B_T, T) \leq W_n^j(A), \quad 1 \leq j \leq J, \quad n = 1, 2, \cdots. \tag{5.29}$$

As a consequence of Theorem 3.1 and Corollary 5.1, we obtain the sensitivity of the waiting times with respect to the partitioning of $B = B_D + B_T$ when the data buffer is finite.

Theorem 5.2 *Assume $B_D \geq 0$, $B_T \geq 1$ and for all $1 \leq j \leq J$ and all $n \geq 1$, $T \geq \sigma_n^j$ a.s. Then for any fixed cell arrival sequence A,*

$$W_n^j(TB, B_D + 1, B_T, T) \leq W_n^j(TB, B_D, B_T + 1, T), \quad 1 \leq j \leq J, \quad n = 1, 2, \cdots \tag{5.30}$$

$$W_n^j(LB, B_D + 1, B_T, T) \leq W_n^j(LB, B_D, B_T + 1, T), \quad 1 \leq j \leq J, \quad n = 1, 2, \cdots \tag{5.31}$$

6 Comparison of Queue Lengths

Since the service discipline is first come first serve, the queue length seen by any customer is increasing in its waiting time. More precisely, we have that for all $1 \leq j \leq J$ and all $n \geq 1$,

$$N_n^j = \begin{cases} 0, & W_n^j = 0; \\ \inf\{m \geq 1 \mid \sum_{i=1}^{m} \sigma_{n-i}^j \geq W_n^j, \ \sum_{i=1}^{m-1} \sigma_{n-i}^j < W_n^j.\}, & W_n^j > 0. \end{cases}$$

Therefore,

Theorem 6.1 *All the assertions of Theorems 5.1 and 5.2, and Corollaries 5.1 and 5.2 are valid with W_n^j replaced by N_n^j.*

We prove now that the queue lengths at all time instants are bounded. We consider the case of single downstream queue $J = 1$.

Theorem 6.2 *Assume $B_D \geq 0$, $B_T \geq 1$ and for all $n \geq 1$, $T \geq \sigma_n^1$ a.s. Then for any fixed cell arrival sequence A,*

$$M_t(TB, B_D, B_T, T) \leq B_T + 1, \qquad t \geq 0, \tag{6.1}$$

$$M_t(LB, B_D, B_T, T) \leq B_T, \qquad t \geq 0. \tag{6.2}$$

Proof. Consider first (6.1). We show by induction on n that relations (6.3) and (6.4) below hold.

$$M_t(TB, B_D, B_T, T) + Q_t^T(TB, B_D, B_T, T) \leq B_T + 1,$$
$$\widehat{g}_{n-1}(TB, B_D, B_T, T) \leq t < \widehat{g}_n(TB, B_D, B_T, T), \tag{6.3}$$

$$M_{\widehat{s_n^-}(TB,B_D,B_T,T)}(TB, B_D, B_T, T) + Q^T_{\widehat{s_n^-}(TB,B_D,B_T,T)}(TB, B_D, B_T, T) < B_T + 1. \tag{6.4}$$

Clearly, they are true for $n \leq B_T$. Assume they hold for some $n \geq B_T$. Then, since there is at most one customer arriving at the downstream queue at time $\widehat{g}_n(TB, B_D, B_T, T)$, the inductive assumptions (cf. (6.3) and (6.4)) implies that

$$M_{\widehat{g}_n(TB,B_D,B_T,T)}(TB, B_D, B_T, T) + Q^T_{\widehat{g}_n(TB,B_D,B_T,T)}(TB, B_D, B_T, T) \leq B_T + 1. \tag{6.5}$$

During the time interval $[\widehat{g}_n(TB, B_D, B_T, T), \widehat{g}_{n+1}(TB, B_D, B_T, T))$, the quantity $M_t(TB, B_D, B_T, T) + Q_t^T(TB, B_D, B_T, T)$ is unchanged when there is an arrival at the downstream queue and is decreased by one each time there is a departure at the downstream queue. Since there is at least one departure during that time interval if $M_{\widehat{g}_n(TB,B_D,B_T,T)}(TB, B_D, B_T, T) \geq 1$, we conclude from (6.5) that relations (6.3) and (6.4) hold for $n + 1$. Therefore, by induction, the relation

$$M_t(TB, B_D, B_T, T) + Q_t^T(TB, B_D, B_T, T) \leq B_T + 1$$

holds for all $t \geq 0$, so that (6.1) holds.

Consider now (6.2). The idea of the proof is similar. We show by induction on n that relations (6.6) and (6.7) below hold.

$$M_t(LB, B_D, B_T, T) + Q_t^T(LB, B_D, B_T, T) \leq B_T,$$
$$\widehat{g}_{n-1}(LB, B_D, B_T, T) \leq t < \widehat{g}_n(LB, B_D, B_T, T), \tag{6.6}$$

$$M_{s_n^-(LB,B_D,B_T,T)}(LB,B_D,B_T,T) + Q^T_{s_n^-(LB,B_D,B_T,T)}(LB,B_D,B_T,T) < B_T. \qquad (6.7)$$

Clearly, they are true for $n \leq B_T$. Assume they hold for some $n \geq B_T$. Then, since there is at most one customer arriving at the downstream queue at time $\hat{g}_n(LB,B_D,B_T,T)$, the inductive assumptions (cf. (6.6) and (6.7)) implies that

$$M_{\hat{g}_n(LB,B_D,B_T,T)}(LB,B_D,B_T,T) + Q^T_{\hat{g}_n(LB,B_D,B_T,T)}(LB,B_D,B_T,T) \leq B_T. \qquad (6.8)$$

During the time interval $[\hat{g}_n(LB,B_D,B_T,T), \hat{g}_{n+1}(LB,B_D,B_T,T))$, the quantity $M_t(LB,B_D,B_T,T)+Q_t^T(LB,B_D,B_T,T)$ is unchanged when there is an arrival at the downstream queue and is decreased by one each time there is a departure at the downstream queue. If

$$M_{\hat{g}_n(LB,B_D,B_T,T)}(LB,B_D,B_T,T) + Q^T_{\hat{g}_n(LB,B_D,B_T,T)}(LB,B_D,B_T,T) < B_T,$$

then, it is easily seen that relations (6.6) and (6.7) hold for $n+1$. If, however,

$$M_{\hat{g}_n(LB,B_D,B_T,T)}(LB,B_D,B_T,T) + Q^T_{\hat{g}_n(LB,B_D,B_T,T)}(LB,B_D,B_T,T) = B_T,$$

then,

- either $M_{\hat{g}_n(LB,B_D,B_T,T)}(LB,B_D,B_T,T) \geq 1$, in which case, there is at least one departure in the downstream queue during the time interval $[\hat{g}_n(LB,B_D,B_T,T), \hat{g}_{n+1}(LB,B_D,B_T,T))$, so that we conclude from (6.8) that relations (6.6) and (6.7) hold for $n+1$;
- or, $Q^T_{\hat{g}_n(LB,B_D,B_T,T)}(LB,B_D,B_T,T) = B_T$, in which case, the next token arrival occurs only T time units after a cell arrives (this cell is immediately transmitted to the downstream queue), so that, again, there is one departure in the downstream queue during the time interval $[\hat{g}_n(LB,B_D,B_T,T), \hat{g}_{n+1}(LB,B_D,B_T,T))$. Hence, relation (6.8) implies that (6.6) and (6.7) hold for $n+1$.

Therefore, by induction, the relation

$$M_t(LB,B_D,B_T,T) + Q_t^T(LB,B_D,B_T,T) \leq B_T$$

holds for all $t \geq 0$, so that (6.2) holds. ∎

7 Comparison of Losses

In this section, we consider the case when the last queue in the downstream tandem queueing network has finite-capacity waiting buffer. We compare the number of losses in the last queue. We will assume that all the queues in the tandem network have deterministic service times: for all $1 \leq j \leq J$ and all $n \geq 1$, $\sigma_n^j = \sigma^j$, $a.s.$ Note however that the results in this section hold when the last queue has service times with increasing failure rate (IFR) distributions, see the remark at the end of this section.

Our comparison results are based on the following lemma concerning the two $\cdot/G/1$ queues $\mathcal{Q}^{(1)}$ and $\mathcal{Q}^{(2)}$ described at the beginning of Section 5. Let $\{L_t^{(j)}\}$ denote the process of the number of losses in $\mathcal{Q}^{(j)}$, $j = 1, 2$.

Lemma 7.1 *For the two $\cdot/G/1$ queues $\mathcal{Q}^{(1)}$ and $\mathcal{Q}^{(2)}$,*

$$L_t^{(1)} \leq L_t^{(2)}, \qquad t \geq 0. \tag{7.1}$$

provided service times are deterministic and the buffer capacity is finite.

Proof. See Appendix A. ∎

Consider first the case that the data buffer of the rate control throttles have infinite-capacity data buffers: $B_D = \infty$.

Theorem 7.1 *If $B_T \geq 1$ and for all $1 \leq j \leq J$, $T \geq \sigma^j$, then for any fixed cell arrival sequence A,*

$$L_t(LB, \infty, B_T, T) \leq L_t(TB, \infty, B_T, T), \tag{7.2}$$

$$L_t(TB, \infty, B_T, T) \leq L_t(LB, \infty, B_T + 1, T), \tag{7.3}$$

$$L_t(TB, \infty, B_T, T) \leq L_t(TB, \infty, B_T, T'), \qquad T' = mT, m \in I\!N_+ \tag{7.4}$$

$$L_t(LB, \infty, B_T, T) \leq L_t(LB, \infty, B_T, T'), \qquad T \geq T' \tag{7.5}$$

$$L_t(TB, \infty, B_T, T) \leq L_t(TB, \infty, B_T', T'), \qquad T \geq T', 1 \leq B_T < B_T' \tag{7.6}$$

for $t \geq 0$.

Proof. The proof is similar to that of Theorem 5.1 using induction and Theorems 4.2, 4.3 and 4.4, together with Lemmas 5.1 and 7.1. The details are left to the interested reader. ∎

Theorem 7.1 implies the following monotonicity of losses with respect to the token buffer size.

Corollary 7.1 *If $B_T \geq 1$ and for all $1 \leq j \leq J$, $T \geq \sigma^j$, then for any fixed cell arrival sequence A,*

$$L_t(TB, \infty, B_T, T) \leq L_t(TB, \infty, B_T + 1, T), \qquad t \geq 0, \tag{7.7}$$

$$L_t(LB, \infty, B_T, T) \leq L_t(LB, \infty, B_T + 1, T), \qquad t \geq 0. \tag{7.8}$$

When the token buffer has infinite capacity, the downstream tandem network is fed by the arrival sequence A. Denote by $L_t(A)$ the number of lost cells by time t in such a case. The following result indicates that the leaky bucket and the token bank flow control schemes reduce losses:

Corollary 7.2 *If $B_T \geq 1$ and for all $1 \leq j \leq J$, $T \geq \sigma^j$, then for any fixed cell arrival sequence A,*

$$L_t(TB, \infty, B_T, T) \leq L_t(A), \qquad t \geq 0, \tag{7.9}$$

$$L_t(LB, \infty, B_T, T) \leq L_t(A), \qquad t \geq 0. \tag{7.10}$$

Consider now the case when the data buffers of the rate control throttles are finite: $B_D < \infty$. As a consequence of Theorem 3.1, together with Corollary 7.1, we obtain the sensitivity of losses with respect to the partitioning of $B = B_D + B_T$ when the data buffer is finite.

Theorem 7.2 *If $B_T \geq 1$ and for all $1 \leq j \leq J$, $T \geq \sigma^j$, then for any fixed cell arrival sequence A,*

$$L_t(TB, B_D + 1, B_T, T) \leq L_t(TB, B_D, B_T + 1, T), \qquad t \geq 0, \tag{7.11}$$

$$L_t(LB, B_D + 1, B_T, T) \leq L_t(LB, B_D, B_T + 1, T), \qquad t \geq 0. \tag{7.12}$$

Remark: Note that when the service times at the downstream queue are independent and identically distributed with an IFR distribution, all the results in this section together with Lemma A.1 in Appendix A hold, with the inequality $\leq$ replaced by the stochastic inequality $\leq_{st}$. A non-negative random variable X has an IFR distribution if $f_X(x)/[1-F_X(x)]$ is an increasing function of x where $f_X(x)$ is the probability density function of X and $F_X(x) = \int_0^x f_X(y)dy$. Two random variables X and Y are comparable by $\leq_{st}$, say $X \leq_{st} Y$, if for all $x \in I\!R$, $P[X > x] \leq P[Y > x]$. The proofs of Lemmas A.1 and 7.1 can be carried out using coupling arguments (see e.g. [11]).

8 Summary

In this paper we studied the effect that a rate-control throttle has on delays incurred by cells belonging to the session being controlled. We modelled the path taken by cells belonging to that session as a tandem queueing network and showed that the cell delays at each switch increase as the token buffer capacity increases and/or the token generation rate increases. In the case that the buffer corresponding to the last switch on the path has limited buffer capacity, we showed that the number of losses also increases as the token buffer capacity and/or the token generation rate increase. We also established comparisons between the two rate-control throttles. Under appropriate assumptions on the cell arrival process, similar results can be obtained for stationary cell delays and cell losses. Examples of the types of assumptions required can be found in [1] and [12].

A Proof of Lemma 7.1

We first prove the following lemma which will be essential in establishing Lemma 7.1.

Lemma A.1 *Consider two single-server queues $G/D/1/c$ with the same arrival sequence, the same deterministic service time, and the same finite capacity c of the waiting buffer (the server has no buffer). The service discipline is first come first serve. Customers that find waiting buffer full are lost. Let L_t (resp. L'_t) be the number of lost customers by time $t \geq 0$ in the first queue referred to as queue Q (resp. in the second queue referred to as queue Q'). Let M_t (resp.*

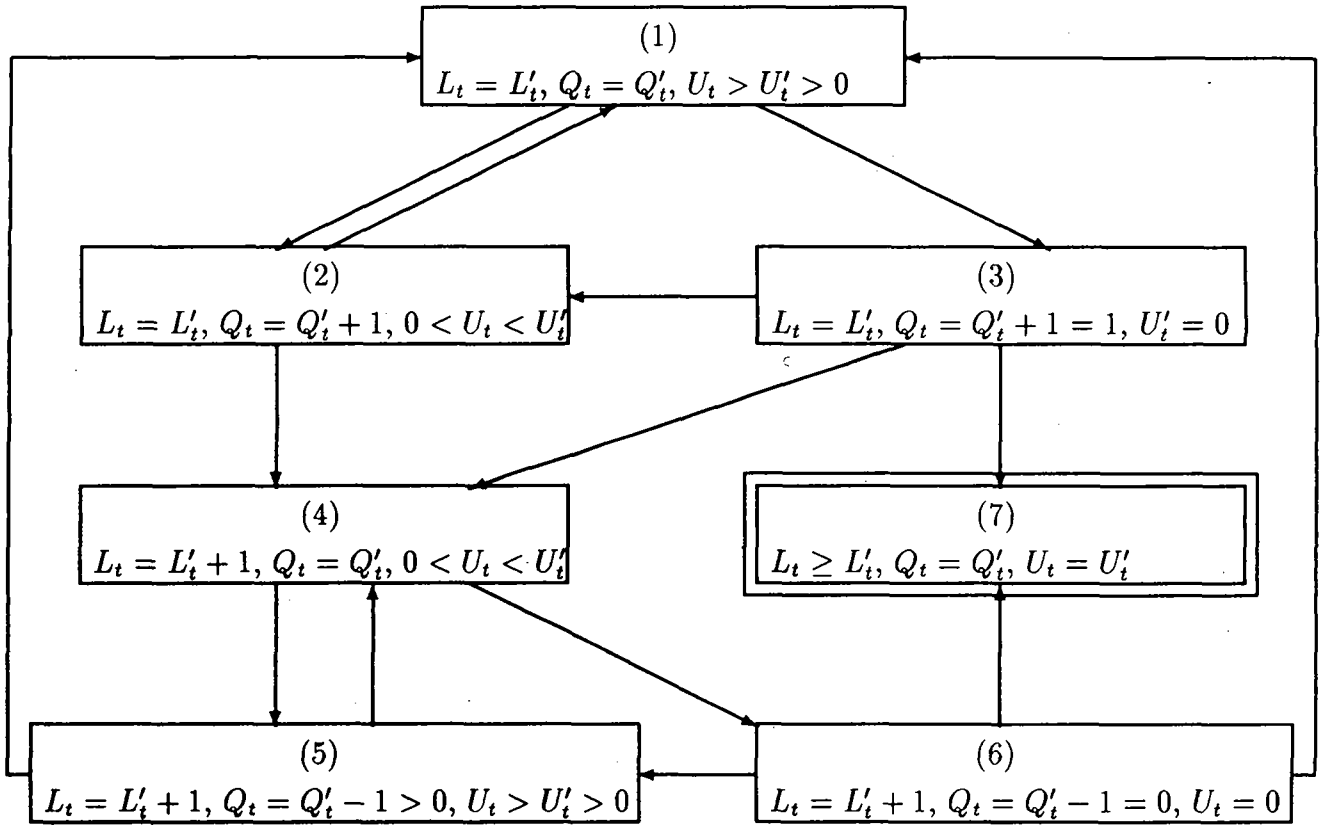

Figure 2: Finite state machine representation of two systems.

M'_t) be the queue length of queue Q (resp. Q') at time $t \geq 0$. Let U_t (resp. U'_t) be the remaining service time of the customer under service (if any) in queue Q (resp. Q') at time $t \geq 0$. The initial conditions of the two systems are such that:

$$L_0 = L'_0, \qquad M_0 = M'_0 \geq 1, \qquad U_0 > U'_0 > 0.$$

Then for all $t \geq 0$,

$$L_t \geq L'_t.$$

Proof. We describe the behavior of the two systems when given the same arrival sequence by a finite state machine (Figure 2) containing seven states which differ according to the relative values of L_t, L'_t, M_t, M'_t, U_t and U'_t.

The systems start in state (1). This state remains in this state whenever a customer arrives (in both systems). When a service completion occurs in Q', the systems transit to state (2) if Q' has waiting customers, and to state (3) if Q' becomes empty.

The systems return to (1) from (2) whenever a service completion occurs in Q. The systems remain in state (2) when a customer arrives (in both systems) resulting in no loss in queue Q (i.e., $M_t < c$), They transit to state (4) otherwise.

The systems transit out of state (3) whenever there is an arrival (in both systems) or there is a service completion in Q). If there is an arrival, the transition is to state (2), provided $M_t < c$, and to (4), provided $M_t = c$. If there is a service completion in queue Q, the transition is to state (7).

No transition occurs from state (4) unless a service completion takes place in Q. In this case the transition is to (5), if Q has waiting customers, and to state (6), if Q' becomes empty.

A transition occurs from state (5) to (4) whenever a service completion takes place in queue Q'. If a customer arrives, no transition occurs if $M_t < c$, or a transition to (1) occurs if $M_t = c$.

If a customer arrives (in both systems) while the systems are in state (6) the systems enter either state (5), provided $M'_t < c$, or state (1), provided $M'_t = c$. If, however, a service completion occurs in system Q', then the systems enter state (7).

State (7) is the absorbing state. Once the systems enter this state, their service completions and their losses synchronize, so that they never transit out of the state.

Since in all these states, we have $L_t \geq L'_t$, the assertion of the lemma is thus established. ∎

Proof of Lemma 7.1

Assume first that $k_0 = \infty$. In addition to the notation previously defined for $Q^{(j)}$, we introduce $U_t^{(j)}$ to be the remaining service time of the customer in service (if any) at time $t \geq 0$.

Define a new queue, $\widehat{Q}$, with the same buffer capacity, say c, as $Q^{(1)}$ and $Q^{(2)}$. Customers arrive to this queue at the same time as to $Q^{(1)}$, at times $\hat{t}_n = t_n^{(1)}$, $n = 1, 2, \ldots$. The customer

service times coincide with those in $Q^{(2)}$ except that the remaining service time of the customer in service (if any) at time $\hat{t}_{n_k}$ is increased to σ if $Q^{(2)}$ is idle, i.e., $M^{(2)}_{\hat{t}^-_{n_k}} = 0$. Let $\widehat{L}_t$ (resp. $\widehat{M}_t$, $\widehat{U}_t$) be the number of losses (resp. queue length, remaining service time of the customer in service) of the queue at time t.

Repetitive applications of Lemma A.1 yields

$$L^{(1)}_t \leq \widehat{L}_t, \qquad t \geq 0. \tag{A.1}$$

We will show by induction that

$$\widehat{U}_{\hat{t}_{n_k}} = U^{(2)}_{\hat{t}_{n_k}}, \qquad k = 1, 2, \cdots, \tag{A.2}$$

$$\widehat{M}_{\hat{t}_{n_k}} \geq M^{(2)}_{\hat{t}_{n_k}}, \qquad k = 1, 2, \cdots, \tag{A.3}$$

$$\widehat{M}_{\hat{t}_{n_k}} + \widehat{L}_{\hat{t}_{n_k}} \leq M^{(2)}_{\hat{t}_{n_k}} + L^{(2)}_{\hat{t}_{n_k}}, \qquad k = 1, 2, \cdots, \tag{A.4}$$

$$\widehat{L}_t \leq L^{(2)}_t, \qquad t \geq 0. \tag{A.5}$$

For $k = 1$, it is clear that $n_1 = 1$ so that $\widehat{U}_{\hat{t}_{n_k}} = U^{(2)}_{\hat{t}_{n_k}} = \sigma$, $\widehat{M}_{\hat{t}_{n_k}} = M^{(2)}_{\hat{t}_{n_k}} = 1$, and $\widehat{L}_{\hat{t}_1} = L^{(2)}_{\hat{t}_1} = 0$. Therefore, relations (A.2), (A.3) and (A.4) hold for $k = 1$, and relation (A.5) holds for all $0 \leq t \leq \hat{t}_{n_1}$.

Assume there is some $k \geq 1$ such that relations (A.2), (A.3) and (A.4) hold for k and relation (A.5) holds for all $0 \leq t \leq \hat{t}_{n_k}$. Consider $k+1$. There are two cases according to whether $n_{k+1} > n_k + 1$ or not.

Case 1: $n_{k+1} = n_k + 1$: By the inductive assumption (cf. (A.3)) we have that for all $t_{n_k} \leq t < t_{n_{k+1}}$, $\widehat{M}_t \geq M^{(2)}_t$. If $M^{(2)}_{t^-_{n_{k+1}}} > 0$, then (cf. (A.2))

$$\widehat{U}_{t_{n_{k+1}}} = U^{(2)}_{t_{n_{k+1}}} = \left(t_{n_{k+1}} - t_{n_k} - U^{(2)}_{t_{n_k}}\right) \bmod \sigma.$$

If, however, $M^{(2)}_{t^-_{n_{k+1}}} = 0$, then, according to the definition of the service times in queue $\widehat{Q}$,

$$\widehat{U}_{t_{n_{k+1}}} = \sigma = U^{(2)}_{t_{n_{k+1}}}.$$

Thus, (A.2) holds for $k+1$.

Since $\widehat{M}_{t_{n_{k+1}}^-} \geq M^{(2)}_{t_{n_{k+1}}^-}$, and since queues $\widehat{Q}$ and Q' have the same buffer capacity, we easily see that $\widehat{M}_{t_{n_{k+1}}} \geq M^{(2)}_{t_{n_{k+1}}}$. Hence, (A.3) holds for $k+1$.

During the time interval $(t_{n_k}, t_{n_{k+1}})$, the inductive assumption (cf. (A.2) and (A.3)) implies that the service completions in queues $Q^{(2)}$ and $\widehat{Q}$ are synchronized unless $Q(2)$ empties. Therefore, the inductive assumption of (A.4) implies that for all $t \in (t_{n_k}, t_{n_{k+1}})$,

$$\widehat{M}_t + \widehat{L}_t \leq M_t^{(2)} + L'_t.$$

At time $t_{n_{k+1}}$, the arrival customer is either accepted or lost, so that $\widehat{M}_t + \widehat{L}_t$ and $M_t^{(2)} + L_t^{(2)}$ are both increased by one. Hence, (A.4) holds for all $k+1$.

It now follows from (A.3) and (A.4) that

$$\widehat{L}_{t_{n_{k+1}}} \leq L^{(2)}_{t_{n_{k+1}}}.$$

As clearly for all $t \in (t_{n_k}, t_{n_{k+1}})$,

$$\widehat{L}_t = \widehat{L}_{t_{n_k}} \leq L^{(2)}_{t_{n_k}} = L_t^{(2)},$$

we conclude that (A.5) holds for all $t \leq t_{n_{k+1}}$.

Case 2: $n_{k+1} > n_k + 1$: Recall from the definition of $\widehat{Q}$, $Q^{(1)}$ and $Q^{(2)}$ that

$$\hat{t}_{n_k+1} - \hat{t}_{n_k} \leq T, \qquad \hat{t}_{n_k+2} - \hat{t}_{n_k+1} = \cdots = \hat{t}_{n_{k+1}-1} - \hat{t}_{n_{k+1}-2} = T, \qquad \hat{t}_{n_{k+1}} - \hat{t}_{n_{k+1}-1} \geq T. \quad \text{(A.6)}$$

Due to the fact that $\sigma \leq T$, we see that during the time interval $(\hat{t}_{n_k}, \hat{t}_{n_{k+1}}]$, only the customer arriving at time $\hat{t}_{n_k+1}$ may be lost in queue $\widehat{Q}$.

If there is no loss in queue $\widehat{Q}$ during the time interval $(\hat{t}_{n_k}, \hat{t}_{n_{k+1}}]$, then, clearly,

$$\widehat{L}_t = \widehat{L}_{\hat{t}_{n_k}} \leq L^{(2)}_{\hat{t}_{n_k}} \leq L_t^{(2)}, \qquad \hat{t}_{n_k} < t \leq \hat{t}_{n_{k+1}}. \quad \text{(A.7)}$$

If, however, this loss does occur in queue $\widehat{Q}$, then, $\hat{t}_{n_k+1} - \hat{t}_{n_k} < \widehat{U}_{\hat{t}_{n_k}}$ and $\widehat{M}_{\hat{t}^-_{n_k+1}} = \widehat{M}_{\hat{t}_{n_k}} = c$. If $M^{(2)}_{\hat{t}_{n_k}} = \widehat{M}_{\hat{t}_{n_k}}$, then, as

$$\hat{t}^{(2)}_{n_k+1} - \hat{t}^{(2)}_{n_k} \leq \hat{t}_{n_k+1} - \hat{t}_{n_k} < \widehat{U}_{\hat{t}_{n_k}} = U^{(2)}_{\hat{t}_{n_k}},$$

we obtain that $M^{(2)}_{\hat{t}^-_{n_k+1}} = M^{(2)}_{\hat{t}_{n_k}} = c$, so that the customer arrving at $\hat{t}^{(2)}_{n_k+1}$ in queue $\mathcal{Q}^{(2)}$ is also lost. Hence,

$$\widehat{L}_t = \begin{cases} \widehat{L}_{\hat{t}_{n_k}} \leq L^{(2)}_{\hat{t}_{n_k}} = L^{(2)}_t, & \hat{t}_{n_k} < t < t^{(2)}_{n_k+1}, \\ \widehat{L}_{\hat{t}_{n_k}} \leq L^{(2)}_{\hat{t}_{n_k}} + 1 \leq L^{(2)}_t, & t^{(2)}_{n_k+1} \leq t < \hat{t}_{n_k+1}, \\ \widehat{L}_{\hat{t}_{n_k}} + 1 \leq L^{(2)}_{\hat{t}_{n_k}} + 1 \leq L^{(2)}_t, & \hat{t}_{n_k+1} \leq t \leq \hat{t}_{n_{k+1}}. \end{cases} \tag{A.8}$$

If $M^{(2)}_{\hat{t}_{n_k}} \neq \widehat{M}_{\hat{t}_{n_k}}$, then, by the inductive assumption (A.3), we have $M^{(2)}_{\hat{t}_{n_k}} < \widehat{M}_{\hat{t}_{n_k}}$. Using further the inductive assumption (A.4), we obtain $\widehat{L}_{\hat{t}_{n_k}} < L^{(2)}_{\hat{t}_{n_k}}$. Therefore,

$$\widehat{L}_t \leq \widehat{L}_{t_{n_k}} + 1 \leq L^{(2)}_{\hat{t}_{n_k}} \leq L^{(2)}_t, \qquad \hat{t}_{n_k} < t \leq \hat{t}_{n_{k+1}}. \tag{A.9}$$

Relations (A.7), (A.8) and (A.9) imply that (A.5) holds for all $t \leq \hat{t}_{n_{k+1}}$.

Let $A^{(2)}(s,t)$, $D^{(2)}(s,t)$ and $L^{(2)}(s,t)$ (resp. $\widehat{A}^{(2)}(s,t)$, $\widehat{D}^{(2)}(s,t)$ and $\widehat{L}^{(2)}(s,t)$) be the number of accepted arrivals, departures and lost arrivals, respectively, in queue $\mathcal{Q}^{(2)}$ (resp. $\widehat{\mathcal{Q}}$) during the time interval $(s,t]$.

We observe from the above arguments that

$$\widehat{M}_{\hat{t}_{n_k}} + \widehat{A}(\hat{t}_{n_k}, \hat{t}_{n_k+1}) \geq M^{(2)}_{\hat{t}_{n_k}} + A^{(2)}(\hat{t}_{n_k}, \hat{t}_{n_k+1}). \tag{A.10}$$

Since the arrivals occur earlier in queue $\mathcal{Q}^{(2)}$ than in queue $\widehat{\mathcal{Q}}$, i.e., $t^{(2)}_n \leq \hat{t}_n$ for all $n \geq 1$, the server in $\widehat{\mathcal{Q}}$ idles during the time interval $(\hat{t}_{n_k}, \hat{t}_{n_{k+1}})$, only if the server in $\mathcal{Q}^{(2)}$ idles. Using further relation (A.6), we conclude that the server in queue $\mathcal{Q}^{(2)}$ idles (resp. $\widehat{\mathcal{Q}}$) during the time interval $(\hat{t}_{n_k}, \hat{t}_{n_{k+1}})$ if and only if $M^{(2)}_{\hat{t}^-_{n_{k+1}}} = 0$ (resp. $\widehat{M}_{\hat{t}^-_{n_{k+1}}} = 0$), and that $\widehat{M}_{\hat{t}^-_{n_{k+1}}} = 0$ implies $M^{(2)}_{\hat{t}^-_{n_{k+1}}} = 0$.

We now distinguish two subcases:

(2.a) $\widehat{M}_{\hat{t}_{n_{k+1}}^-} = 0$ (so that $M^{(2)}_{\hat{t}_{n_{k+1}}^-} = 0$).

Then,

$$\widehat{M}_{\hat{t}_{n_{k+1}}} = M^{(2)}_{\hat{t}_{n_{k+1}}} = 1,$$

$$\widehat{U}_{\hat{t}_{n_{k+1}}} = U^{(2)}_{\hat{t}_{n_{k+1}}} = \sigma,$$

and, owing to (A.5),

$$\widehat{M}_{\hat{t}_{n_{k+1}}} + \hat{L}_{\hat{t}_{n_{k+1}}} \leq M^{(2)}_{\hat{t}_{n_{k+1}}} + L^{(2)}_{\hat{t}_{n_{k+1}}}.$$

Hence, relations (A.2), (A.3) and (A.4) hold for $k+1$.

(2.b) $\widehat{M}_{\hat{t}_{n_{k+1}}^-} > 0$.

It is easily seen that

$$A^{(2)}(\hat{t}_{n_k}, \hat{t}_{n_k+1}) + L^{(2)}(\hat{t}_{n_k}, \hat{t}_{n_k+1}) = \hat{A}(\hat{t}_{n_k}, \hat{t}_{n_k+1}) + \hat{L}(\hat{t}_{n_k}, \hat{t}_{n_k+1}) = n_{k+1} - n_k.$$

Moreover,

$$\begin{aligned} D^{(2)}(\hat{t}_{n_k}, \hat{t}_{n_k+1}) &\leq \left\lfloor (\hat{t}_{n_{k+1}} - \hat{t}_{n_k} - \widehat{U}_{\hat{t}_{n_k}})/\sigma \right\rfloor + 1, \\ &= \left\lfloor (\hat{t}_{n_{k+1}} - \hat{t}_{n_k} - U^{(2)}_{\hat{t}_{n_k}})/\sigma \right\rfloor + 1, \\ &= \hat{D}(\hat{t}_{n_k}, \hat{t}_{n_k+1}). \end{aligned}$$

Hence

$$\begin{aligned} \widehat{M}_{\hat{t}_{n_{k+1}}} + \hat{L}_{\hat{t}_{n_{k+1}}} &= \widehat{M}_{\hat{t}_{n_k}} + \hat{L}_{\hat{t}_{n_k}} + (n_{k+1} - n_k) - \hat{D}(\hat{t}_{n_k}, \hat{t}_{n_k+1}) \\ &\leq M^{(2)}_{\hat{t}_{n_k}} + L^{(2)}_{\hat{t}_{n_k}} + (n_{k+1} - n_k) - D^{(2)}(\hat{t}_{n_k}, \hat{t}_{n_k+1}), \\ &= M^{(2)}_{\hat{t}_{n_{k+1}}} + L^{(2)}_{\hat{t}_{n_{k+1}}}. \end{aligned}$$

Hence, relation (A.4) holds for $k+1$.

If $M^{(2)}_{\hat{t}^-_{n_{k+1}}} > 0$, then

$$D^{(2)}(\hat{t}_{n_k}, \hat{t}_{n_k+1}) = \hat{D}(\hat{t}_{n_k}, \hat{t}_{n_k+1}) = \left\lfloor (\hat{t}_{n_{k+1}} - \hat{t}_{n_k} - U^{(2)}_{\hat{t}_{n_k}})/\sigma \right\rfloor + 1,$$

so that it follows from (A.10) that

$$\begin{aligned} \widehat{M}_{\hat{t}_{n_{k+1}}} &= \widehat{M}_{\hat{t}_{n_k}} + \hat{A}(\hat{t}_{n_k}, \hat{t}_{n_k+1}) - \hat{D}(\hat{t}_{n_k}, \hat{t}_{n_k+1}) \\ &\geq M^{(2)}_{\hat{t}_{n_k}} + A^{(2)}(\hat{t}_{n_k}, \hat{t}_{n_k+1}) - D^{(2)}(\hat{t}_{n_k}, \hat{t}_{n_k+1}) \\ &= M^{(2)}_{\hat{t}_{n_{k+1}}}. \end{aligned}$$

Furthermore,

$$\hat{U}_{\hat{t}_{n_{k+1}}} = U^{(2)}_{\hat{t}_{n_{k+1}}} = \left(\hat{t}_{n_{k+1}} - \hat{t}_{n_k} - U^{(2)}_{\hat{t}_{n_k}}\right) \bmod \sigma.$$

Hence, relations (A.2) and (A.3) hold for $k+1$.

If, however, $M^{(2)}_{\hat{t}^-_{n_{k+1}}} = 0$, then

$$\widehat{M}_{\hat{t}_{n_{k+1}}} \geq 1 = M^{(2)}_{\hat{t}_{n_{k+1}}},$$

and, according to the definition of the service times in queue $\hat{\mathcal{Q}}$,

$$\hat{U}_{\hat{t}_{n_{k+1}}} = U^{(2)}_{\hat{t}_{n_{k+1}}} = \sigma.$$

Thus, again, relations (A.2) and (A.3) hold for $k+1$.

This completes the inductive proof. If $k_0 < \infty$, then, due to the fact that $\hat{t}_{n_{k_0}+1} - \hat{t}_{n_{k_0}} \leq T$ and $\hat{t}_i - \hat{t}_{i-1} = T$ for all $i \geq n_{k_0} + 2$, similar arguments can be used to show that relation (A.5) holds for all $t \leq \hat{t}_{n_{k_0}+1}$. Since there is no loss in queue $\hat{\mathcal{Q}}$ after time $\hat{t}_{n_{k_0}+1}$, we conclude that (A.5) remains true for all $t > \hat{t}_{n_{k_0}+1}$.

Combining relations (A.1) and (A.5) implies that relation (7.1) holds for all $t \geq 0$. The proof is thus completed. ∎

References

[1] V. Anantharam, P. Konstantopoulos, "Burst Reduction Properties of the Leaky Bucket Flow Control Scheme in ATM Networks", manuscript.

[2] A. W. Berger, "Performance Analysis of a Rate-Control Throttle where Tokens and Jobs Queue", *IEEE J-SAC*, Vol. 9, pp. 165–170, 1991.

[3] A. W. Berger, W. Whitt, "The Impact of a Job Buffer in a Token-bank Rate-control Throttle" , to appear in *Stochastic Models*, 1992.

[4] K. C. Budka, *Sample Path Analysis of Flow Control Schemes for Packet Networks.* PhD thesis, Harvard University, Division of Applied Sciences, July 1991.

[5] R.L. Cruz, H.-N. Liu "Non-recursive identities for tandem queueing networks", preprint.

[6] A. Demers, S. Keshav, and S. Shenker, "Analysis and simulation of a fair queueing algorithm", *Internetworking:Research and Experience*, **1**, 3 - 26, May 1990.

[7] S.J. Golestani, "A stop-and-go framework for congestion management", *Proc. 1990 SIGCOMM*, 8-18, June 1990.

[8] C.R. Kalmanek, H. Kanakia, and S. Keshav, "Rate controlled servers for very high-speed networks", *Proc. Globecom'90*, Dec. 1990.

[9] L. Kuang, "On the variance reduction property of buffered leaky bucket", SRC TR 91-90, University of Maryland, USA, 1991.

[10] L. Kuang, "Monotonicity properties of the leaky bucket", SRC TR 92-27, University of Maryland, USA, 1992.

[11] Z. Liu, D. Towsley, "Optimality of the round robin routing policy". COINS Technical Report, TR 92-55, 1992. To appear *Journal of Applied Probability.*

[12] Z. Liu, D. Towsley, "Burst reduction properties of rate-based flow control schemes: departure process", To appear in *Annals of Operations Research*, Special Issue on Methodologies for Performance Analysis of High Speed Networks.

[13] S. Low, P. Varaiya, "Burstiness bounds for some burst reducing servers", *Proc. INFOCOM'93*, 2–8, 1993.

[14] E.P. Rathgeb, "Modeling and performance comparison of policing mechanisms for ATM networks", *IEEE J. Sel. areas Comm.*, **9**, 3, pp. 325-334, 1991.

[15] M. Sidi, W. Liu, I. Cidon, I. Gopal, "Congestion control through input rate regulation", *Proc. GLOBECOM'89*, 1989.

[16] J. Turner, "New Directions in communications (or which way to the information age)", **24**, 8 –15, 1986.

Unité de Recherche INRIA Sophia Antipolis
2004, route des Lucioles - B.P. 93 - 06902 SOPHIA ANTIPOLIS Cedex (France)

Unité de Recherche INRIA Lorraine Technopôle de Nancy-Brabois - Campus Scientifique
615, rue du Jardin Botanique - B.P. 101 - 54602 VILLERS LES NANCY Cedex (France)
Unité de Recherche INRIA Rennes IRISA, Campus Universitaire de Beaulieu 35042 RENNES Cedex (France)
Unité de Recherche INRIA Rhône-Alpes 46, avenue Félix Viallet - 38031 GRENOBLE Cedex (France)
Unité de Recherche INRIA Rocquencourt Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 LE CHESNAY Cedex (France)

EDITEUR
INRIA - Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 LE CHESNAY Cedex (France)

ISSN 0249 - 6399